THE FORMATION OF CONDENSED CORRELATION TABLES WHEN THE NUMBER OF COMBINATIONS IS LARGE

DR. J. ARTHUR HARRIS

CARNEGIE INSTITUTION OF WASHINGTON

AFTER the principles of any method of research are laid down by those who have the genius or the good fortune to make fundamentally new contributions, there always remains much to be done in the refinement, simplification, or adaptation of methods to render them most practically applicable in the routine of investigation. This is especially true in the modern higher statistics, where, at the very best, the labor is excessive.

One of the most onerous of the statistical processes is the determination of correlation in cases in which each individual measurement must be weighted by comparison with a series of others. In an earlier number of this journal[1] a method was described for the rapid formation of the heavy intra-class and inter-class[2] correlation and contingency surfaces by the use of a machine permitting simultaneous multiplication and summation. Methods of dealing with such correlations without the formation of tables will be published later. But abstract formulæ in the hands of inexperienced calculators are apt to lead to erroneous constants, which in the absence of the original data can never be corrected. Again, the validity of the correlation coefficient as a measure of interdependence depends largely upon linearity of regression. Hence, tables should be given whenever possible. The purpose of this note is to show how, in the case of relationships

[1] "On the Formation of Correlation and Contingency Tables when the Number of Combinations is Large," AMER. NAT., Vol. 45, pp. 566–571, 1911.

[2] These terms will be clear from their context in this note; they will be more precisely defined later.

involving a very large number of combinations, the chief advantages of the correlation (but not the contingency) surface may be even more easily realized than in the method already described.

By condensed correlation tables are to be understood those giving the (weighted) frequencies for a first character x and the first (and where necessary also the second) rough moment about 0 as origin of the associated array of the y character.[3] From such a table[4] r may be quickly obtained[5] and the means of arrays calculated for linearity of regression tests.

In principle, the formation of these reduced tables is very simple. Let $x, y, z, \cdots$, be measures on the individuals of the same or associated classes. Let there be $n, p, q, \cdots$, of these individuals. Then if $n, p, q, \cdots, \Sigma(x')$, $\Sigma(y')$, $\Sigma(z')$, $\cdots$, $\Sigma(x'^2)$, $\Sigma(y'^2)$, $\Sigma(z'^2)$, $\cdots$ (where Σ indicates a summation within the class and the dashes indicate that the measures are to be regarded as deviations from 0) be again summed for each of the component measures, seriated by grades, the four columns—grade of "first individual," weighted frequency, and the two rough moments about 0 for associated individuals—thus secured for each character either constitute the desired table or one from which it may be easily derived.

The arithmetical routine will be determined largely by the nature of the records. Roughly, two cases are possible: $n, p, q, \cdots$, are small, m is small or large; $n, p, q, \cdots$, are large, m is small,[6] m being the number of classes or groups of classes.

Suppose $n, p, q, \cdots$, small, say 4–20. The best method

[3] In direct intra-class correlations x and y are measures of the same kind; in cross intra-class correlations they are different; in inter-class relationships they may be the same or different.

[4] For example, Table X of *Biometrika*, Vol. 8, p. 61, 1911, or Table II derived from Table I of the AMER. NAT., Vol. 44, p. 695, 1910.

[5] See "The Arithmetic of the Product Moment of Calculating the Coefficient of Correlation," AMER. NAT., Vol. 44, pp. 693–699, 1910.

[6] Cases where both the numbers within the class and the number of classes are large are very rare because of the great labor required in making the observations.

is to write the values of the first character under consideration—designated for convenience as the subject—down the side of a separate sheet for each class. Opposite each entry is then written n, $\Sigma(x')$ and $\Sigma(x'^2)$, p, $\Sigma(y')$ and $\Sigma(y'^2)$, q, $\Sigma(z')$ and $\Sigma(z'^2)$ and so on, according to the relationships desired. Thus, the measure used as the subject and the number and summed first and second powers of deviation of the individuals of the relative array may be for the same or different characters or classes, depending on whether direct or cross, intra-class or inter-class correlation is to be computed. In any case, the number and moments are only once determined for each class—their repeated entry on the sheet is merely rapid clerical work.

This done, the sheets are clipped into strips by subject entries, the strips seriated according to the subject, and the class numbers and moments summed for each grade on the machine.

For inter-class correlations, the resulting table is correct, embracing as it does, say, $S(pq)$ entries. For intra-class relationships, say for x, the entries are too high by $S(n)$, $S(x')$ and $S(x'^2)$ since it comprises $S(n^2)$ entries when only $Sn(n—1)$ are desired. Hence, the actual frequency for each subject grade must be subtracted from the weighted frequency, and the products of the actual frequency by the grade and by the square of the grade must be deducted from the first and second summed moment column, respectively.

When the number of individuals per class, n, p, q, is large (*e. g.*, 25 or over) another procedure is desirable. The classes of the subject character are seriated (in transverse rows) in a table of vertical columns captioned by the grades. Opposite each row is entered n, $\Sigma(x')$ and $\Sigma(x'^2)$, p, $\Sigma(y')$ and $\Sigma(y'^2)$, q, $\Sigma(z')$ and $\Sigma(z'^2)$, $\cdots$, for all characters to be correlated. The associated (weighted) values for each subject grade are quickly gathered by multiplying up and summing simultaneously the fre-

quencies in each column of the subject seriations by the opposed entries in the relative (number and summation) columns. Again, the result is the desired table or one from which it may be derived.

Illustrations will make the methods most clear. Table I shows the frequencies for the different grades of radial asymmetry[7] of quinquilocular fruits gathered from 34 individuals of *Hibiscus Syriacus* in the Missouri Botanical Garden in the fall of 1907. Table II gives the seriations for the locular composition[8] of the same fruits. The last two columns of Table I and the next to the last two of Table II give the first two summations for each individual.[9]

[7] The radial asymmetry is the standard deviation of the number of ovules per locule about the mean number of ovules per fruit. See *Biometrika*, Vol. 7, pp. 476–479, 1910, for full discussion.

At the head of this table the coefficients of asymmetry are for condensation given to only two places. In all the calculations, however, they have been used to six places. Their values and their true squares as used in the calculations are:

Asymmetry a	a^2
.000000	.00
.400000	.16
.489897	.24
.632455	.40
.748331	.56
.800000	.64
.894427	.80
.979795	.96
1.019803	1.04
1.095445	1.20
1.166190	1.36
1.200000	1.44
1.264911	1.60
1.356466	1.84
1.600000	2.56

[8] Expressed here simply as the number of locules per fruit with an "odd" number of ovules. *Cf. Biometrika*, Vol. 7, pp. 483–487, 1910.

[9] The last two columns of Table II give the summations of Table I for convenience in determining the cross intra-class tables. When the cross intra-class tables are to be formed with asymmetry as the subject the $\Sigma(c')$ and $\Sigma(c'^2)$ column may be added to Table I. Here it is omitted for convenience in publication.

TABLE I

SERIATIONS AND SUMMATIONS OF RADIAL ASYMMETRY BY INDIVIDUALS

Tree	Radial Asymmetry—Standard Deviations of Individual Fruits															N	$\Sigma(a')$	$\Sigma(a'^2)$
	.00	.40	.48	.63	.74	.80	.89	.97	1.01	1.09	1.16	1.20	1.26	1.35	1.60			
1	22	25	20	5	14	9	2	—	1	1	—	—	—	—	—	99	44.540951	28.24
2	37	29	19	5	5	2	—	1	1	—	—	—	—	—	—	99	31.411571	17.28
3	46	34	13	2	1	2	—	1	—	—	—	—	—	—	—	99	24.561697	12.16
4	66	29	4	1	—	2	—	—	—	—	—	—	—	—	—	102	15.792043	7.28
5	11	25	33	8	11	4	3	—	4	—	1	—	—	—	—	100	50.586565	31.76
6	14	22	38	13	8	5	2	—	2	—	1	1	—	—	—	106	51.819299	32.00
7	46	30	16	3	3	2	—	—	—	—	—	—	—	—	—	100	25.580710	12.80
8	13	33	29	6	11	3	2	1	1	—	—	—	—	—	—	99	45.621836	26.32
9	13	27	32	4	12	10	1	2	—	—	—	—	1	—	—	102	50.105424	31.04
10	26	40	17	6	4	6	1	1	—	—	1	—	—	—	1	103	40.556715	24.64
11	16	30	25	11	10	4	4	—	—	—	1	—	—	—	—	101	46.631638	27.92
12	8	21	34	14	8	12	1	—	—	—	1	—	—	—	—	99	51.558133	31.44
13	34	19	24	6	3	9	1	1	—	—	—	—	—	—	—	97	34.471473	20.40
14	59	33	4	1	—	—	—	—	—	—	—	—	—	—	—	97	15.792043	6.64
15	13	21	30	10	14	1	2	3	1	1	—	—	—	2	—	98	50.254513	33.44
16	42	28	18	6	1	3	—	1	—	—	—	—	—	—	—	99	27.941002	14.64
17	33	26	22	7	8	4	—	—	—	—	—	—	—	—	—	100	34.791567	19.28
18	50	24	18	2	2	—	—	2	1	—	—	—	—	—	—	99	24.159111	13.04
19	63	18	11	2	2	3	—	—	—	—	—	—	—	—	—	99	17.750439	9.36
20	72	20	6	—	—	1	—	1	—	—	—	—	—	—	—	100	12.719177	6.24
21	41	29	16	4	3	4	—	—	2	—	1	—	—	—	—	100	30.618961	17.76
22	42	21	25	4	5	2	—	1	—	—	—	—	—	—	—	100	29.498695	16.00
23	34	31	19	5	4	4	1	2	—	—	—	—	—	—	—	100	33.917659	19.04
24	31	16	28	1	10	6	3	3	1	—	—	—	—	—	—	99	39.675350	25.44
25	15	33	32	4	6	10	—	1	—	—	—	—	—	—	—	101	44.876305	25.28
26	28	26	21	8	7	5	2	—	1	—	—	—	—	—	—	98	37.794451	22.16
27	56	30	8	3	1	2	—	—	—	—	—	—	—	—	—	100	20.164872	9.76
28	38	34	20	2	3	2	1	—	—	—	—	—	—	—	—	100	29.402270	14.80
29	4	25	35	5	11	11	2	2	—	—	—	1	—	—	—	96	52.288755	32.56
30	40	25	14	5	5	5	3	1	—	—	—	—	—	—	—	98	31.425564	18.72
31	11	27	34	7	9	8	—	2	—	—	1	1	—	—	—	100	49.344442	30.16
32	28	31	12	13	9	2	2	3	1	—	1	—	—	—	—	102	41.749890	26.24
33	28	21	20	4	6	11	4	3	—	—	1	1	—	—	—	99	42.901029	29.04
34	7	34	29	7	13	8	1	1	1	—	—	1	—	—	—	102	52.456526	31.84

TABLE II

SERIATIONS AND SUMMATIONS FOR LOCULAR COMPOSITION BY INDIVIDUALS

Tree	Locular Composition—Number of "Odd" Locules per Fruit						N	$\Sigma(c')$	$\Sigma(c'^2)$	$\Sigma(a')$	$\Sigma(a'^2)$
	0	1	2	3	4	5					
1	25	23	19	22	9	1	99	168	466	44.540951	28.24
2	36	25	18	12	6	2	99	131	351	31.411571	17.28
3	46	36	11	5	1	–	99	77	141	24.561697	12.16
4	67	30	3	2	—	–	102	42	60	15.792043	7.28
5	10	19	24	33	12	2	100	224	654	50.586565	31.76
6	13	18	38	24	11	2	106	220	612	51.819299	32.00
7	44	31	9	13	1	2	100	102	250	25.580710	12.80
8	10	21	25	22	17	4	99	225	691	45.621836	26.32
9	15	27	31	17	9	3	102	191	523	50.105424	31.04
10	31	37	18	10	7	–	103	131	311	40.556715	24.64
11	13	30	25	22	7	4	101	194	540	46.631638	27.92
12	8	27	28	29	6	1	99	199	521	51.558133	31.44
13	35	24	27	6	3	2	97	118	284	34.471473	20.40
14	59	33	4	1	—	–	97	44	58	15.792043	6.64
15	9	19	34	24	8	4	98	211	599	50.254513	33.44
16	42	31	14	11	1	–	99	96	202	27.941002	14.64
17	31	22	23	14	7	3	100	153	427	34.791567	19.28
18	50	26	18	5	—	–	99	77	143	24.159111	13.04
19	66	18	8	7	—	–	99	55	113	17.750439	9.36
20	72	21	5	1	1	–	100	38	66	12.719177	6.24
21	42	27	20	6	4	1	100	106	250	30.618961	17.76
22	41	18	19	15	4	3	100	132	368	29.498695	16.00
23	35	33	14	14	3	1	100	120	288	33.917659	19.04
24	32	19	23	17	6	2	99	150	410	39.675350	25.44
25	17	36	28	14	5	1	101	159	379	44.876305	25.28
26	26	22	23	14	10	3	98	165	475	37.794451	22.16
27	57	31	10	2	—	–	100	57	89	20.164872	9.76
28	38	30	16	9	7	–	100	117	287	29.402270	14.80
29	8	27	31	20	10	–	96	189	491	52.288755	32.56
30	44	27	15	9	3	–	98	96	216	31.425564	18.72
31	11	14	37	14	19	5	100	231	717	49.344442	30.16
32	28	33	24	12	4	1	102	138	326	41.749890	26.24
33	30	26	17	14	5	7	99	157	475	42.901029	29.04
34	7	25	29	21	19	1	102	227	659	52.456526	31.84

TABLE III

LOCULAR COMPOSITION

		0	1	2	3	4	5	Totals
Radial Asymmetry	.000000	1,038	—	—	—	—	49	1,087
	.400000	—	730	—	—	187	—	917
	.489897	—	—	420	306	—	—	726
	.632455	—	—	73	111	—	—	184
	.748331	—	—	179	30	—	—	209
	.800000	45	101	—	—	12	4	162
	.894427	—	37	—	—	1	—	38
	.979795	14	12	—	—	5	2	33
	1.019803	—	—	6	11	—	—	17
	1.095445	—	—	1	1	—	—	2
	1.166190	—	—	8	1	—	—	9
	1.200000	—	5	—	—	—	—	5
	1.264911	—	1	—	—	—	—	1
	1.356466	—	—	1	1	—	—	2
	1.600000	1	—	—	—	—	—	1
	Totals,	1,098	886	688	461	205	55	3,393

TABLE IV

Asymmetry and Asymmetry

Asymmetry	Gross Values			Values to be Deducted			Working Table		
	n	Total a'	Total a'^2	n	Total a'	Total a'^2	n	Total a'	Total a'^2
.000000	108,324	32,152.8593	18,159.68	1,087	000.0000	000.00	107,237	32,152.8593	18,159.68
.400000	91,608	33,189.7416	19,387.04	917	366.8000	146.72	90,691	32,822.9416	19,240.32
.489897	72,526	29,615.2738	17,726.56	726	355.6652	174.24	71,800	29,259.6085	17,552.32
.632455	18,427	7,668.6159	4,616.56	184	116.3717	73.60	18,243	7,552.2441	5,427.28
.748331	20,879	9,099.5847	5,544.32	209	156.4012	117.04	20,670	8,943.1835	5,427.28
.800000	16,162	6,787.1336	4,103.04	162	129.6000	103.68	6,000	6,657.5336	3,999.36
.894427	3,790	1,663.8007	1,030.72	38	33.9882	30.40	3,752	1,629.8125	1,000.32
.979795	3,285	1,318.9137	807.68	33	32.3332	31.68	3,252	1,286.5805	776.00
1.019803	1,707	734.8870	450.56	17	17.3367	17.68	1,690	717.5503	432.88
1.095445	197	94,7955	61.68	2	2.1909	2.40	195	92.6046	59.28
1.166190	910	405.7667	250.96	9	10.4957	12.24	901	395.2710	238.72
1.200000	503	248.8101	155.60	5	6.0000	7.20	498	242.8100	148.40
1.264911	102	50.1054	31.04	1	1.2649	1.60	101	48.8405	29.44
1.356466	196	100.5090	66.88	2	2.7129	3.68	194	97.7961	63.20
1.600000	103	40.5567	24.64	1	1.6000	2.56	102	38.9567	22.08
	338,719	123,171.3535	72,416.96	3,393	1,232.7607	724.72	335,326	121,938.5929	71,692.24

TABLE V

Locular Composition and Asymmetry

Locular Composition	Gross Values			Values to be Deducted			Working Table		
	n	Total a'	Total a'^2	n	Total a'	Total a'^2	n	Total a'	Total a'^2
0 odd	109,418	32,490.3521	18,369.20	1,098	51.3171	44.80	108,320	32,439.0350	18,324.40
1 odd	88,448	31,547.4534	18,410.64	886	424.9163	231.36	87,562	31,122.5371	18,179.28
2 odd	68,767	28,239.0889	16,957.92	688	403.7775	250.40	68,079	27,835.3114	16,707.52
3 odd	46,075	19,480.3436	11,759.60	461	257.3969	150.48	45,614	19,222.9468	11,609.12
4 odd	20,521	9,073.6754	5,490.16	205	90.1934	43.20	20,316	8,983.4820	5,446.96
5 odd	5,490	2,340.4401	1,429.44	55	5.1596	4.48	5,435	2,335.2805	1,424.96
	338,719	123,171.3535	72,416.96	3,393	1,232.7607	724.72	335,326	121,938.5928	71,692.24

From I and II, the machine quickly compiles four working tables—a direct intra-class for asymmetry, *a*, and another for locular composition, *c*, and two cross intra-class tables.[10] The columns under "gross values" in

TABLE VI

ASYMMETRY AND LOCULAR COMPOSITION

A	Gross Values			Values to be Deducted			Working Table		
	n	Total c'	Total c'^2	n	Total c'	Total c'^2	n	Total c'	Total c'^2
.00	108,324	117,335	288,699	1,087	245	1,225	107,237	117,090	287,474
.40	91,608	127,391	332,887	917	1,478	3,722	90,691	125,913	329,165
.48	72,526	117,550	318,254	726	1,758	4,434	71,800	115,792	313,820
.63	18,427	30,358	81,952	184	479	1,291	18,243	29,879	80,661
.74	20,879	36,577	100,993	209	448	986	20,670	36,129	100,007
.80	16,162	26,180	70,560	162	169	393	16,000	26,011	70,167
.89	3,790	6,549	18,093	38	41	53	3,752	6,508	18,040
.97	3,285	5,014	13,422	33	42	142	3,252	4,972	13,280
1.01	1,707	3,040	8,460	17	45	123	1,690	2,995	8,337
1.09	197	379	1,065	2	5	13	195	374	1,052
1.16	910	1,600	4,406	9	19	41	901	1,581	4,365
1.20	503	1,024	2,954	5	5	5	498	1,019	2,949
1.26	102	191	523	1	1	1	101	190	522
1.35	196	422	1,198	2	5	13	194	417	1,185
1.60	103	131	311	1	0	0	102	131	311
	338,719	473,741	1,243,777	3,393	4,740	12,442	335,326	469,001	1,231,335

TABLE VII

LOCULAR COMPOSITION AND LOCULAR COMPOSITION

Loc. Comp.	Gross Values			Values to be Deducted			Working Table		
	n	Total c'	Total c'^2	n	Total c'	Total c'^2	n	Total c'	Total c'^2
0	109,418	117,758	288,576	1,098	0,000	0,000	108,320	117,758	288,576
1	88,448	118,815	305,713	886	886	886	87,562	117,929	304,827
2	68,767	111,775	302,475	688	1,376	2,752	68,079	110,399	299,723
3	46,075	78,151	213,853	461	1,383	4,149	45,614	76,768	209,704
4	20,521	37,538	105,540	205	820	3,280	20,316	36,718	102,260
5	5,490	9,704	27,620	55	275	1,375	5,435	9,429	26,245
	338,719	473,741	1,243,777	3,393	4,740	12,442	335,326	469,001	1,231,335

Tables IV–VII give the results. These contain, since $p=q$, a total $S(p^2)=S(q^2)=S(pq)$ entries, whereas in the direct intra-class relationships $S[p(p-1)]=S[q(q-1)]$, and in the cross intra-class $S[p(q-1)=S[q(p-1)]$ are desired.

[10] One for the relationship between radial asymmetry and locular composition, the other for the correlation between locular composition and radial asymmetry. Of course, both give the same end result, and only one need be found unless the linearity of both regressions is to be tested.

From these gross values must be deducted, therefore, the actual frequency for each grade of the subject and the product of the frequency by the first and second power of the grade in the case of direct intra-class correlation, or the frequency of the grade and the sum of the first and second powers of the values of the relative character in the same fruit in the cross intra-class correlation. Data for these are given in the table showing the correlation for asymmetry and locular composition of the same fruit, Table III. The second set of three columns in Tables IV–VII gives the quantities so calculated from Table III to be deducted. The final three columns are in each case the working tables.

The first and second moments for the (weighted) population A and σ are given by the totals of the two final columns. Or those for the subject character may be calculated (and a check for the accuracy of the totals secured) from the grade of the subject and the weighted frequency column.[11]

From our working tables, indicating by S a summation from our final tables, we determine by the methods of Amer. Nat., Vol. 45, pp. 693–699, 1910, these values:

For Asymmetry

$$S(a') = 121{,}938.5928, \quad A_a = .363642,$$
$$S(a'^2) = 71{,}692.2400, \quad \sigma_a = .285593.$$

For Locular Composition

$$S(c') = 469{,}001, \quad A_c = 1.398642,$$
$$S(c'^2) = 1{,}231{,}335, \quad \sigma_c = 1.309906.$$

For Asymmetry and Locular Composition

Table IV, $S(a_1'a_2') = 48{,}818.9505$, $r = .1637$,
Table VI, $S(a_1'c_2') = 192{,}072.3309$,[12] $r = .1716$,
Table V, $S(c_1'a_2') = 192{,}072.3308$,[12] $r = .1716$,

[11] Of course in practise, the second population moment may be calculated by $S[(n-1)\Sigma(x'^2)]$, $S[(p-1)\Sigma(y'^2]$, $S[(q-1)\Sigma(z'^2)]$, . . . , thus obviating the labor of forming the third columns, which are included here for completeness of illustration merely.

[12] The difference of .0001 is due to the necessity of lopping off the last two places of the six decimals in the asymmetry coefficient in the one case while they can be retained in the other. Of course, it is of no practical significance.

Table VII, $S(c_1'c_2') = 763{,}048.0000$, $r = .1861$.

While primarily illustrations of method, these results, if they are substantiated by further work, seem to me of considerable biological interest. They show not only that individuals of *H. Syriacus* differ in the radial asymmetry and in the locular composition of their fruits, but that when an individual bears fruits above the average asymmetry, it also produces fruits above the average in number of "odd" locules. Apparently, this cross correlation is as high as either of the direct correlations.

Two biological interpretations are possible. (*a*) The production of radially symmetrical ovaries and those with a high number of odd locules depends upon the same morphogenetic tendencies of the primordia,[13] which give rise to the fruit. (*b*) There is in *Hibiscus* an intra-individual selective elimination similar to that demonstrated in *Staphylea*,[14] the intensity of which differs from individual to individual in such a way as to bring about (statistical) correlation for characters originally uncorrelated.

The discussion of these points falls outside the scope of the present note where the data serve merely as a random illustration of a very rapid method of carrying out the routine of a widely applicable statistical process.

COLD SPRING HARBOR,
April 25, 1912

[13] In the individual fruit radial asymmetry and locular composition are necessarily associated (*cf. Biometrika*, Vol. 7, pp. 491–493, 1910). In *Staphylea*, correlations of $r = .22$ to $r = .33$ have been noted. Table III above gives $r = .527$ for asymmetry and locular composition of the same fruit.

Probably in all these relationships regression is not linear, and the correlations must be interpreted with caution.

[14] *Biometrika*, Vol. 7, pp. 452–504, 1910; *Science*, N. S., Vol. 32, pp. 519–528, 1910; *Zeitschr. f. Ind. Abst. u. Vererbungsl.*, Vol. 5, pp. 273–288, 1911; *Pop. Sci. Mo.*, Vol. 78, pp. 534–537, 1911.